Whole blood gene expression profiling of neonates with confirmed bacterial sepsis

Genomics Data 3 (2015) 41–48

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Whole blood gene expression profiling of neonates with confirmed bacterial sepsis

Paul Dickinson [a,b,*,1], Claire L. Smith [c,1], Thorsten Forster [a,b], Marie Craigon [a], Alan J. Ross [a], Mizan R. Khondoker [a,2], Alasdair Ivens [d,3], David J. Lynn [e,4], Judith Orme [c], Allan Jackson [c], Paul Lacaze [a], Katie L. Flanagan [f,5], Benjamin J. Stenson [c], Peter Ghazal [a,b,*]

[a] Division of Infection and Pathway Medicine, Edinburgh Infectious Diseases, University of Edinburgh, Edinburgh EH16 4SB, UK
[b] SynthSys—Synthetic and Systems Biology, University of Edinburgh, Edinburgh EH9 3JD, UK
[c] Neonatal Unit, Simpson Centre for Reproductive Health, Royal Infirmary of Edinburgh, Edinburgh EH16 4SA, UK
[d] Fios Genomics Ltd., Edinburgh BioQuater, Edinburgh EH16 4SB, UK
[e] Animal & Bioscience Research Department, AGRIC, Teagasc, Grange, Dunsany, Co. Meath, Ireland
[f] MRC Research Laboratories, Atlantic Boulevard, PO Box 273, Fajara, Gambia

ARTICLE INFO

Article history:
Received 31 October 2014
Accepted 6 November 2014
Available online 15 November 2014

Keywords:
Neonatal sepsis
Whole blood
Gene expression profiling
Microarray

ABSTRACT

Neonatal infection remains a primary cause of infant morbidity and mortality worldwide and yet our understanding of how human neonates respond to infection remains incomplete. Changes in host gene expression in response to infection may occur in any part of the body, with the continuous interaction between blood and tissues allowing blood cells to act as biosensors for the changes. In this study we have used whole blood transcriptome profiling to systematically identify signatures and the pathway biology underlying the pathogenesis of neonatal infection. Blood samples were collected from neonates at the first clinical signs of suspected sepsis alongside age matched healthy control subjects. Here we report a detailed description of the study design, including clinical data collected, experimental methods used and data analysis workflows and which correspond with data in Gene Expression Omnibus (GEO) data sets (GSE25504). Our data set has allowed identification of a patient invariant 52-gene classifier that predicts bacterial infection with high accuracy and lays the foundation for advancing diagnostic, prognostic and therapeutic strategies for neonatal sepsis.

Specifications	
Organism/cell line/tissue	Homo sapiens/whole blood
Sex	Male and female
Sequencer or array type	Illumina HT-12V3.0 Whole Human Genome microarray, CodeLink 55K Whole Human Genome microarray, Affymetrix U219 Whole Human Genome microarray and Affymetrix HG U133 Plus 2.0 Whole Human Genome microarray
Data format	Raw data (Tab delimited text files of background subtracted signals and .CEL files)
Experimental factors	Blood culture or cerebrospinal fluid positive bacterial sepsis vs. healthy control whole blood samples and culture negative suspected infected samples
Experimental features	A case–control gene expression profiling study of whole blood taken from neonates at the first clinical sign of sepsis and control healthy neonates. Study includes training and replication sets for blood culture positive samples and clinical evaluation set of blood culture negative sepsis cases. Results compared blood culture or cerebrospinal fluid positive septic neonates, blood culture negative septic neonates and healthy control neonates. Prior power calculations were based on Healthy Edinburgh neonates using the CodeLink platform and Gambian infants (9 months of age) were used for further refinement of power calculations using Illumina HT-12 platform.

* Corresponding authors at: Division of Pathway Medicine, Edinburgh Infectious Diseases, University of Edinburgh, Edinburgh EH16 4SB, UK.
E-mail addresses: paul.dickinson@ed.ac.uk (P. Dickinson), p.ghazal@ed.ac.uk (P. Ghazal).
[1] These authors contributed equally to this work.
[2] Present address: Department of Biostatistics, Institute of Psychiatry and NIHR Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust, King's College, London, UK.
[3] Present address: Centre for Infection Immunity and Evolution, King's Buildings, University of Edinburgh, Edinburgh, UK.
[4] Present address: EMBL Australia Laboratory, South Australian Health and Medical Research Institute, North Terrace, Adelaide, South Australia 5000, Australia.
[5] Present address: Department of Immunology, Monash University, Commercial Road, Prahran, Melbourne, Victoria 3181, Australia.

http://dx.doi.org/10.1016/j.gdata.2014.11.003

(*continued*)

Specifications	
Consent	Written informed consent was obtained from parents of all enrolled infants in accordance with approval granted by the Lothian Research Ethics Committee for blood samples for RNA isolation obtained at the first time of clinical signs of suspected sepsis (reference 05/s1103/3). Samples obtained from The Gambia conformed to MRC policy regarding ethical research in children and were approved by the local scientific coordinating committee (SCC), the Joint Gambia Government/MRC Ethics Committee and by the London School of Hygiene and Tropical Medicine Ethics Committee (reference SCC1085 Pilot Study 1 (L2008.63))
Sample source location	Edinburgh, UK and The Gambia

Direct link to deposited data

Deposited data are available here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25504.

Experimental design, materials and methods

Patient demographics and experimental design

The study was conducted in the Neonatal Unit, Royal Infirmary of Edinburgh and the Division of Pathway Medicine, University of Edinburgh. The patient demographics, microbial organisms isolated and reasons for blood sampling in controls for all patient sets are shown in Table 1. Infants having blood cultures taken to investigate suspected infection (Table 1B) and "well" control infants having blood taken for other clinical reasons (Table 1C) were studied. Samples taken from patients with suspected clinical infection that proved to have microbiological evidence of infection from a usually sterile body site were identified and formed the infected group. Full clinical assessment for early and late symptoms and signs of sepsis followed criteria for neonatal sepsis taken from data as detailed in Table 2, with the blood culture test used as the 'gold standard' for diagnosis of sepsis. Five infants had samples included from more than one episode of infection. To meet with laboratory regulations, samples that could be considered 'high risk' were excluded. Infants were not included in the study if the mother was known to be positive for hepatitis B, HIV or hepatitis C viruses. In cases where the mother was known to have a history of drug misuse and had not had antenatal screening for blood-borne viruses, the infants were also excluded. Other exclusion criteria were infants who did not require clinical blood samples and infants for whom extra blood sampling might be of particular risk, for example, infants with an underlying disorder causing anemia. Before embarking on this study we had previously performed a power calculation using the CodeLink chip platform [1] on neonatal samples but we also performed a power calculation using the Illumina chip platform, on an independent set of 30 infant samples at 9 months of age, before vaccination. This showed that the study design has 90% power to detect a twofold change in expression with an a of 1% (false discovery rate (FDR) corrected), for more than 99% of 35,177 gene probes present on the array [2]. A schematic of patient recruitment and sample processing workflow for the samples processed for the training, replication and validation arm of the study is shown in Fig. 1.

Sample collection and RNA extraction

For RNA isolation, blood (500 μl–1 ml) was immediately injected into PAXgene™ blood RNA tubes (PreAnalytiX BD/QIAgen) and mixed

Table 1
Patient demographics of samples used, microorganisms identified from infected patients and reasons for blood sampling in controls.

Patient demographics of samples used

Sample set	Training set		Platform test set		Validation test set	
Infection status	Infected ($n = 28$)	Control ($n = 35$)	Infected ($n = 18$)	Control ($n = 24$)	Infected ($n = 16$)	Control ($n = 10$)
Male	15 (54%)	22 (63%)	10 (56%)	15 (63%)	10 (63%)	9 (90%)
Gestation completed at birth (week): range (mean)	24–38 (28.5)	26–42 (37.9)	24–38 (28.8)	26–42 (37.3)	23–40 (28.3)	24–41 (31)
Gestation completed at sampling (week): range (mean)	26–39 (31.1)	31–44 (39.4)	26–39 (30.8)	31–44 (39.1)	25–41 (33.8)	29–42 (34.9)
Birthweight (g): range (mean)	430–3380 (1126)	650–4570 (3080)	430–3380 (1236)	650–4350 (2941)	635–3160 (1134)	800–4220 (1932)

Microorganisms identified from infected patients

Organism	Training set	Platform test set	Validation test set
Coagulase negative staphylococcus	15	8	7
Enterococcus species	4	3	1
Group B Streptococcus	2	2	1
Klebsiella species	2	1	2
Candida albicans and Klebsiella species	1	1	–
Escherichia Coli	1	1	1
Enterobacter cloacae	1	1	–
Pseudomonas aeruginosa	1	1	1
CMV	1	–	–
Listeria monocytogenes	–	–	1
Serratia marcescens	–	–	2

Reasons for blood sampling in controls

Reason	Training set	Platform test set	Validation test set
Screening test: maternal thyroid disease	17	9	–
Bilirubin check due to jaundice	5	4	1
"Routine" neonatal screening (preterms)	5	4	6
Electrolyte check: previous deranged Na	3	3	–
Screening test: pigmented scrotum	2	1	3
Blood count check: Coomb's positive	1	1	–
Screening test: newborn bloodspot	1	1	–
Neonatal encephalopathy	1	1	–

A. Patient demographics of samples used. Patient sample details are shown displaying the demographics of the population studied. B. Microorganisms identified from infected patients. Organisms detected for each infected infant are shown — these samples were taken at, or within 6 h of, the time of clinical suspicion of infection. C. Reasons for blood sampling in controls. The reasons for clinical blood sampling in the control group are shown — all of the screening tests in these infants were normal. Table 1 was adapted from Supplementary Table 3 of Smith et al. 2014 [2] by permission from Macmillan Publishers Ltd: Nature Communications [2], copyright (2014).

Table 2
Clinical details of patient samples used in the study. Table 2 was adapted from Supplementary Data 4 of Smith et al. 2014 [2] by permission from Macmillan Publishers Ltd: Nature Communications [2], copyright (2014).

Sample no.	Microarray Platform(s)	Study Category	Same Baby	Gestation at birth (completed weeks)	Postnatal age at sample (completed days)	Birthweight (g)	Gender	Ethnicity	Organism	Site	Other	Early or late onset infection	Sample in relation to blood culture (hours)	Duration of Antibiotics (days)
Inf009	Illumina	Infected		28	33	720	Girl	caucasian	Coagulase negative staphylococcus	Blood		late	0	14
Inf012	CodeLink, AffyU133	Infected		34	6	2150	Girl	caucasian	Enterovirus	Csf		late	0	10
Inf032	CodeLink, AffyU133	Infected		40	8	3180	Girl	caucasian	Listeriamonocytogenes	Csf		late	0	14
Inf047	Illumina, CodeLink	Infected		29	12	1380	Boy	caucasian	Candida	Blood	Also Klebsiella in peritoneal fluid	late	0	7
Inf075	Illumina, CodeLink	Infected	70	26	74	890	Girl	caucasian	Cytomegalovirus	Blood, urine		late	0	3
Inf082	Illumina, CodeLink	Infected		29	8	1230	Boy	caucasian	Enterobacter cloacae			late	0	14
Inf083	Illumina, CodeLink	Infected		29	36	1405	Girl	not stated	Enterococcus faecium	Blood		late	0	10
Inf084	Illumina, CodeLink	Infected	116	28	10	720	Boy	caucasian	Coagulase negative staphylococcus	Blood	2 cultures	late	0	14
Inf089	Illumina, CodeLink	Infected		27	9	850	Boy	not stated	Coagulase negative staphylococcus	Blood		late	0	3
Inf091	Illumina, CodeLink	Infected		38	1	3900	Girl	caucasian	Group B streptococcus	Blood		early	0	5
Inf102	Illumina, CodeLink	Infected	103	27	31	785	Girl	not stated	Pseudomonas aeuruginosa	Blood		late	0	14
Inf107	CodeLink	Infected	108,110	23	8	655	Boy	not stated	Coagulase negative staphylococcus	Blood	2 cultures	late	0	7
Inf111	Illumina	Infected	115	28	14	1200	boy	caucasian	Coagulase negative staphylococcus		2 cultures	late	0	9
Inf112	Illumina, CodeLink	Infected	124,129,130	24	10	660	Girl	Oriental	Group B streptococcus	Blood		late	0	11
Inf114	Illumina, CodeLink	Infected	99	26	9	955	Girl	caucasian	Coagulase negative staphylococcus	Blood		late	0	11
Inf116	Illumina	Infected*	84	28	14	720	boy	caucasian	Coagulase negative staphylococcus	Abscess	Blood culture next day positive for coagulase negative staphlococcus	late	0	14
Inf118	CodeLink	Infected	120,122	29	59	940	Boy	caucasian	Coagulase negative staphylococcus	Blood		late	0	10
Inf119	Illumina	Infected	125	28	16	983	girl	not stated	Coagulase negative staphylococcus	Blood		late	0	18
Inf125	Illumina, CodeLink	Infected	119	28	12	983	Girl	not stated	Coagulase negative staphylococcus	Blood		late	0	18
Inf132	Illumina, CodeLink	Infected	131,145	27	16	1100	Boy	not stated	Enterococcus faecalis	Blood		late	0	21
Inf133	Illumina	Infected		27	4	1140	boy	not stated	Coagulase negative staphylococcus	Blood		late	0	5
Inf137	CodeLink	Infected		29	43	1150	Boy	caucasian	Coagulase negative staphylococcus	Blood		late	0	18
Inf138	Illumina, CodeLink	Infected		25	16	870	Girl	caucasian	Coagulase negative staphylococcus	Blood		late	0	10
Inf145	Illumina	Infected	131,132	27	12	1100	boy	not stated	Enterococcus faecalis	Blood		late	0	21
Inf149	Illumina	Infected	148, 151, 155	27	15	840	girl	caucasian	Klebsiella	Csf		late	0	25
Inf152	Illumina, CodeLink	Infected		28	21	450	Boy	caucasian	Coagulase negative staphylococcus	Blood		late	0	15
Inf155	CodeLink	Infected	148,149,151	27	10	840	Girl	caucasian	Klebsiella pneumoniae	Blood		late	0	25
Inf157	Illumina	Infected	158,164,167	28	24	820	girl	caucasian	Coagulase negative staphylococcus	Blood		late	8	16
Inf159	Illumina, CodeLink	Infected		30	9	1335	Boy	caucasian	Enterococcus	Blood		late	0	11
Inf161	CodeLink	Infected		28	9	1090	Boy	caucasian	Klebsiella oxytoca	Blood and csf		late	0	23
Inf162	Illumina, CodeLink	Infected	156	27	6	1130	Boy	asian	Coagulase negative staphylococcus	Blood	2 cultures	late	0	5
Inf164	Illumina, CodeLink	Infected	157,158,167	28	26	820	Girl	caucasian	Klebsiella oxytoca	Blood		late	0	16
Inf191	Illumina, CodeLink	Infected		37	0	2970	Boy	caucasian	Escherichia coli	BlcOd		early	0	7
Inf198	Illumina, CodeLink	Infected	175,185,203	32	8	1070	Boy	caucasian	Coagulasenegative staphylococcus	Blood		late	0	10
Inf203	Illumina	Infected	175,185,198	32	35	1070	boy	caucasian	Coagulase negative staphylococcus	Blood		late	0	6
Inf216	AffyU219	Infected		25	66	810	girl	not stated	Enterococcus	Blood		late	0	16
Inf226	AffyU219	Infected		29	29	1385	boy	not stated	Enterococcus	Blood		late	0	7
Inf239	AffyU219	Infected		27	5	680	boy	not stated	Pseudomonas aeuruginosa	Blood		late	0	2 days (until death)
Inf262	AffyU219	Infected		24	111	690	girl	caucasian	Group B streptococcus	Blood		late	0	13
Inf271	AffyU219	Infected	267	29	48	1040	boy	not stated	Coagulase negative staphylococcus	Blood	Klebsiella 4 days previously in blood, candida 2 days later	late	0	10
Inf275	AffyU219	Infected		28	77	1255	boy	not stated	Coagulase negative staphylococcus	Blood		late	0	7
Inf287	AffyU219	Infected		26	62	880	girl	not stated	Escherichia coli	Blood		late	0	13
Inf297	AffyU219	Infected	283, 286, 299	30	39	1490	boy	not stated	Serratia marcescens	Blood		late	0	13
Inf299	AffyU219	Infected	283, 296, 297	30	44	1490	boy	not stated	Serratia marcescens	Blood		late	0	13
NEC253	AffyU219	Infected		30	44	1390	girl	not stated	Coagulase negative staphylococcus	Blood		late	13	10
NEC283	AffyU219	Infected	297, 299, 286	30	12	1490	boy	not stated	Candida	Peritoneal fluid	Enterobacter blood 3 days earlier	late	11 hours post op	11
Sus002	CodeLink	Possible		41	1	3480	Girl	asian					0	3
Sus005	CodeLink	Possible		25	68	750	Boy	caucasian					0	2
Sus008	CodeLink	Possible		33	0	1370	Girl	caucasian					0	5
Sus010	CodeLink	Possible		38	0	3660	Boy	caucasian					0	5
Sus011	CodeLink	Possible		35	0	2510	Girl	caucasian					0	5
Sus013	CodeLink	Possible		40	0	3530	Boy	Mixed					0	2
Sus019	CodeLink	Possible		29	0	1510	Girl	Caucasian					0	3
Sus020	CodeLink	Possible		34	0	1930	Girl	Caucasian					0	3
Sus024	CodeLink	Possible		34	0	2170	Boy	Caucasian					0	2
Sus025	CodeLink	Possible		34	0	2210	Boy	Caucasian					0	2
Sus026	CodeLink	Possible		35	0	2725	Boy	Caucasian					0	5
Sus027	CodeLink	Possible		41	0	3615	Boy	Asian					0	5
Sus033	CodeLink	Possible		35	0	1840	Boy	Caucasian					0	5
Sus034	CodeLink	Possible		29	0	1490	Girl	Caucasian					0	5
Sus035	CodeLink	Possible		41	1	4860	Boy	Caucasian					0	4
Sus036	CodeLink	Possible		41	0	3360	Girl	Caucasian					0	0
Sus037	CodeLink	Possible		26	32	700	Boy	Caucasian					0	0
Sus038	CodeLink	Possible		29	0	1130	Boy	Caucasian					0	0
Sus039	CodeLink	Possible		29	0	1345	Boy	Caucasian					0	0
Sus041	CodeLink	Possible		29	4	1750	Boy	Caucasian					0	2
Sus044	CodeLink	Possible		36	0	2550	Boy	Caucasian					0	2
Sus053	CodeLink	Possible		35	0	2230	Girl	Hispanic					0	5
Sus054	CodeLink	Possible		37	0	3080	Girl	Caucasian					0	2
Sus055	CodeLink	Possible		41	1	4520	Girl	Caucasian					0	5
Sus057	CodeLink	Possible		37	0	2985	Girl	Afrocaribbean					0	2
Sus059	CodeLink	Possible		41	1	3600	Boy	Not stated					12 hours after	5
Sus061	CodeLink	Possible		30	9	1210	Girl	Not stated					0	10
Sus064	CodeLink	Possible		26	4	970	Girl	Not stated					0	12
Sus068	CodeLink	Possible		40	1	3680	Boy	Not stated					0	0
Sus074	CodeLink	Possible		34	0	2220	Girl	Not stated					0	5
Vir228	AffyU219	Infected		31	24	1920	Boy	Not stated	Rhinovirus	Nasopharyngeal secretions		late	n/a	n/a
Vir269	AffyU219	Infected		35	5	1290	boy	Not stated	Cytomegalovirus	Blood, urine		late	4 days	16 days antiviral therapy
Vir278	AffyU219	Infected	279	24	58	855	boy	Not stated	Rhinovirus	Nasopharyngeal secretions		late	0	0

(continued on next page)

Table 2 (*continued*)

Respiratory distress	Apnoeas, increased oxygen requirement or increased	Bradycardias	Reduced perfusion, hypotension or fluid	Temperature	Metabolic acidosis	Feed intolerance or abdominal concerns	Abnormal tone or irritability	Lethargy	Jaundice	Poor colour or looks unwell	Other	Abnormal lab parameters	Death	Timing of death (days post sample)	Haemoglobin g/l	White cell count	Neutrophil count	Platelets	pH	Sugar	Expert opinion on likelihood of infection in "possibles"	Other diagnoses
X	X	0	0	X	0	X	0	0	0	0		Neutrophilia, low platelets	0		115	16.4	15.03	14	7.29	3.7		
X	0	0	X	X	0	0	0	X	X	X	0	Low platelets	0		184	18.1	5.44	7	7.26	3.44		
0	0	0	0	X	0	0	X	X	X	0	Red groin	Elevated white cell count	0		168	29.4	23.81	320	7.40	5.9		
X	0	0	0	X	0	X	0	0	0	X	Known NEC, central line	Low platelets	X	7	91	9.9	6.34	53	7.28	4.2		
0	X	X	0	0	0	0	0	0	0	X			0		81	5.8	1.18	250				
0	X	X	0	X	0	0	0	0	0	0	Quiet, intubated the next day	Low platelets	0		143	16.2	12.01	0				
0	X	X	0	0	0	0	0	0	0	0		Low platelets	0		89	9.8	3.6	31				
0	0	X	0	X	0	0	0	X	0	0		Low platelets	0		98	13.7	4.8	40		2.36		
0	X	X	X	X	0	X	0	X	X	0	Loose stool	Neutrophilia	0		83	23.9	20.08	453	7.24	6.3		
0	0	0	0	X	0	X	X	0	0	X	Quiet		0		173	10	7.8	200	7.38	3.87		
0	0	0	0	0	X	X	0	X	0	0	Central line, blood sugar instability	Low platelets	X	>30	122	6.9	3.86	N/A	7.20	2.3		
0	0	0	X	0	0	X	0	0	0	0	Hyperglycaemia	Low platelets, high white cell count	X	7	118	31	24	88	7.14	10.4		
0	X	X	X	0	0	X	0	X	0	X		Neutrophilia	0		112	24	22	205	7.22	5.7		
0	X	X	0	X	0	X	0	X	X	X	Hyperglycaemia	High white cell count, neutrophilia	0		109	49.7	44.7	317	7.18	14.9		
0	0	X	0	0	0	0	0	X	0	X			0		130	13	6	287	7.25	8.46		
0	0	0	0	0	0	X	0	0	0	0	Red hot fluctuant swelling forearm	Low platelets	0		111	15.1	6.8	101				
X	X	X	0	0	0	0	0	X	0	X		Low platelets	0		88	3.1	2	108	7.26			
0	X	X	0	X	0	X	0	X	0	X		Low platelets	0		109	17.1	4.8	78	7.20	3.33		
0	X	X	0	0	0	X	0	X	0	X		High white cell count	0		85	14.1	10.86	324	7.29	3.82		
0	X	X	0	0	0	X	0	0	0	0		Low platelets	0		94	14	7.4	60				
0	X	X	0	0	0	0	0	0	X	0	Hyperglycaemia		0		138	7.5	6.6	241	7.26	9.9		
0	X	X	X	X	0	X	0	X	0	X		Low platelets	0		126	16.4	13.6	79				
0	X	X	0	X	0	0	0	X	0	0	Quiet		0		108	13.9	8.3	177	7.37	10.4		
X	X	X	0	0	0	X	0	X	0	0		Low platelets	0		131	6.3	3.8	66				
0	X	X	X	X	0	X	0	X	0	X	Quiet	Low platelets	0		116	7.6	3.2	10	7.22	2.4		
0	X	0	X	0	0	0	0	0	0	0	Trisomy 21	Low platelets, hypoglycaemia	X	12	158	18.3	13.4	92	7.19	2		
X	X	X	0	0	0	X	0	X	0	0			0		143	9	5.2	224	7.28	5.1		
0	0	0	0	0	0	X	0	X	0	0	Surgical wound abscess/granuloma, central line	Low platelets	0		148	5.6	2.2	104	7.37			
X	0	X	0	X	0	0	0	0	X	0		Elevated white cell count, bordeline platelets	0		171	22.7	17.3	142	7.29			
X	X	X	0	0	0	X	0	X	X	X		Low platelets	0		117	7	2.7	47	7.19	6.8		
0	X	X	0	X	0	0	0	X	0	0			0		103	11.1	7.5	237	7.25	11		
X	X	X	X	X	0	0	0	X	0	0	Quiet, previous nec	Low platelets	0		127	10.9	9.2	84	7.21			
X	0	0	0	0	0	0	0	X	0	0		Neutrophilia	0		189	16	14.12	223	7.39	6.3		
X	X	X	X	X	0	0	0	X	0	X	Quiet, tachycardic		0		116	6.8	4.73	287	7.2	4.5		
X	X	X	X	X	0	0	0	X	0	X	Quiet	Low platelets	0		124	7.4	3.94	67	7.38	3.8		
0	X	0	0	0	0	X	0	0	0	0	Glucose instability	Low platelets	0		131	11.8	5.23	15	7.39	3.1		
0	X	0	X	X	0	0	0	0	0	0	Central line		0									
0	X	0	0	0	0	X	0	0	0	X	Central line	Low platelets	X	2	94	11.7	5.5	33		11.5		
X	X	0	X	X	X	0	0	0	0	0	Raised lactate, stoma	Bordeline platelets	0		114	6.6	3.8	125	7.18	5.4		
0	X	0	0	0	0	0	0	0	0	0	Extravasation	Low platelets	0		113	10	3.4	85	7.38	4.2		
0	0	0	0	X	0	0	0	X	0	0	Central line, previous NEC		0		102	19	13.66	196				
0	X	0	X	X	0	X	X	0	0	X			0		110	7.9	3.9	231	7.38	4.7		
0	X	X	X	X	0	0	0	0	0	0		Low white cells, low platelets	0		110	2.7	1.6	104	7.31			
0	X	X	X	X	0	X	0	X	0	X	Central line, blood sugar instability	Low platelets	0		85	2.8	1.8	55	7.23	4.8		
0	0	0	0	0	0	X	0	0	0	0	Rash, sore	Neutrophilia	0		127	27.5	21.1	339				Necrotising enterocolitis
0	0	0	0	0	0	0	0	0	0	0	<12 hours post-op NEC	Low platelets	0		165	16.1	12.2	62	7.43	4		Post-op for necrotising enterocolitis
0	0	0	0	X	0	0	X	0	0	0	Rash		0		156	15.9	11.28	248	7.43		Low	
X	X	X	0	0	0	0	0	0	0	X	Rash, oedema		0		105	6.2	2.7	190	7.25	5.7	High	Patent ductus arteriosus ligated 4 days later
X	X	0	0	0	0	0	0	0	0	0	IUGR		0		193	10.4	3.49	208			Medium	
X	X	X	0	0	0	0	X	X	0	X			0		164	14.3	9.91	199	7.26	4.7	High	
X	0	0	X	0	0	0	X	0	0	0	Sticky eye	Neutrophilia	0		174	29.6	23.29	200	7.26	2	Medium	
X	0	0	0	0	0	0	X	0	0	0			0		181	17.2	9.59	244	7.34	2.3	Low	
0	0	0	0	0	0	0	0	0	0	0			0		174	15.7	8.16	241	7.27	1.9	Low	
0	0	0	0	0	0	0	X	0	0	0	No respiratory effort at birth		0		188	6	2.78	243	7.36	1.47	Low	
X	0	0	0	0	0	0	0	0	0	0			0		186	11.1	2.92	274	7.31	3.64	Low	
X	0	0	0	0	0	0	0	0	0	0	Rash		0		183	12.5	5.08	282	7.22		Low	
X	X	0	0	0	0	0	0	0	0	0	Rash		0		163	14.2	7.83	246	7.12	2.1	Medium	
0	0	0	0	0	0	0	X	X	0	0	Poor suck		0		184	21.8	11.99	306	7.19	3.3	Medium	
X	0	0	0	0	0	0	0	0	0	0	Rash		0		162	11.5	10.22	278	7.27	3.5	Medium	
X	0	0	0	0	0	0	0	0	0	0			0		210	9.3	9.3	241	7.27	0.3	Medium	
X	0	0	0	0	0	0	X	0	0	0			0		196	20.1	12.66	293			Low	
0	0	0	0	X	0	0	0	X	0	0			0		185	20	13.53	208		3.9	Low	
X	X	X	0	0	0	0	0	0	0	X	Quiet		0		96	8.6	2.06	171	7.26	4.2	Low	
0	0	0	0	0	0	0	0	0	0	0	Plethoric	Neutropenia	0		182	6.2	0.97	202	7.35		Low	
0	0	0	0	0	0	0	0	0	0	0	Plethoric	Neutropenia	0		195	5.4	1.43	214	7.34	1.6	Low	
0	0	0	0	0	0	0	0	0	0	X	Hypertension, 2 days post op diaphragmatic hernia		0		119	10.6	6.7	198	7.19	4.4	Medium	Post-op diaphragmatic hernia
X	0	0	0	0	0	0	0	0	0	0	Hypoglycaemia		0		179	10	2.62	286		1.17	Low	
X	X	0	0	X	0	0	0	X	0	X			0		147	6.4	3.72	261	7.16	5.2	High	
X	0	0	0	X	0	0	0	0	0	0	Poor feeding		0		173	21.6	16.2	259	7.31	4.4	Medium	
X	0	0	0	0	0	0	0	0	0	0	Rash		0		231	7.7	5.11	164	7.39	3	Low	
X	0	0	0	0	0	0	0	0	0	0	Starry eyed		0		201	18.4	11.91	189	7.35	3.6	Low	
X	0	0	0	0	0	X	0	X	X	0	Pale, fine creps in chest, spleen tipable	Neutrophilia	0		201	21.6	17.05	310			High	
0	0	0	0	X	0	X	0	0	0	0	?NEC, quiet	High white counr, neutrophilia	0		155	42.4	36.46	392	7.37	4.3	High	
0	X	X	0	0	0	0	0	X	0	0	Pale	Neutrophilia	0		112	29	20.33	288	7.09	6.5	High	
0	0	0	0	X	0	0	0	0	0	0	Offensive liquor		0		179	10.9	4.78	235			Low	
X	0	0	0	0	0	0	0	0	0	0			0		182	16.5	9.05	337	7.19	2	Medium	
0	0	0	0	0	0	0	0	0	0	0			0		86	8.5	2.7	410	7.36	3.9		
0	0	0	0	0	0	0	0	0	X	0	Rash, IUGR, brain cyst, splenomegally	Low platelets	0		179	4.9	2.21	33		3.3		
0	0	0	0	X	0	0	0	0	0	0	Tachycardic, nasal secretions		0		98	10.7	2.5	472	7.34	5.4		

by inversion. Samples were then frozen at −20 °C until RNA extraction occurred as described previously [1]. RNA was quantified and A_{260}:A_{280} ratios generated using a ThermoSpectronic NanoDrop™1000 spectrophotometer. RNA quality was assessed qualitatively by examining the electropherogram and quantitatively from the RNA integrity number (RIN) generated by an Agilent 2100 Bioanalyser.

Gene expression profiling using microarrays

Microarray study design

This study was designed as a prospective case–control study to biologically and computationally infer a set of genes acting as a reliable classifier for bacterial sepsis in neonates. This design entails a main data set to identify genes and train a classification algorithm and is referred to as 'training set'. Subsequent validation of the trained classifier requires independent data sets referred to as 'test sets' (distinctions between them outlined below). Based on earlier power calculations, the training set (Illumina HT-12 v3 platform) was established with 27 patient samples with a confirmed blood culture-positive test for sepsis (bacterial infected cases) and 35 age-matched controls (it also contains one cytomegalovirus-infected case that was not used for classification), all sub-selected by sample quality from the full study population. For assessing reproducibility of our gene classifier with a different assay platform, we examined a subset of 42 of these samples using the CodeLink gene expression platform (comprising 18 bacterial infected and 24 control samples) named in this study as 'platform test set'. Subsequently, for independent clinical evaluation, the 52-gene set classifier was applied to a further 29 new and independent samples (comprising 16 bacterial infected, 3 viral infected and 10 control samples) named in this study as 'validation test set' which were analyzed using the CodeLink, Affymetrix HG-U133 Plus 2.0 and Affymetrix U219 gene expression platforms. The classifier was then used on 30 suspected infected samples and classification of samples into infected and non-infected cases compared against an 'expert' clinical classification.

RNA labeling and hybridization

For Illumina HT-12 v3 arrays total RNA was converted to double-stranded cDNA, followed by an in vitro transcription amplification step to generate labeled cRNA, using the Ambion Illumina TotalPrep-96 RNA Amplification Kit. The cRNA was quantified by A260 measurement using a NanoDrop™1000 spectrophotometer. The cRNA was normalized and hybridized onto the Illumina HT-12 v3 arrays overnight (16 h) at 58 °C. The unhybridized and non-specifically hybridized cRNA was washed away. The arrays are stained with Cy3-Streptavidin to bind to the analytical probes that have hybridized to the array. Arrays were scanned using an Illumina IScan scanner and fluorescence emissions were recorded in high-resolution images. The intensities of the

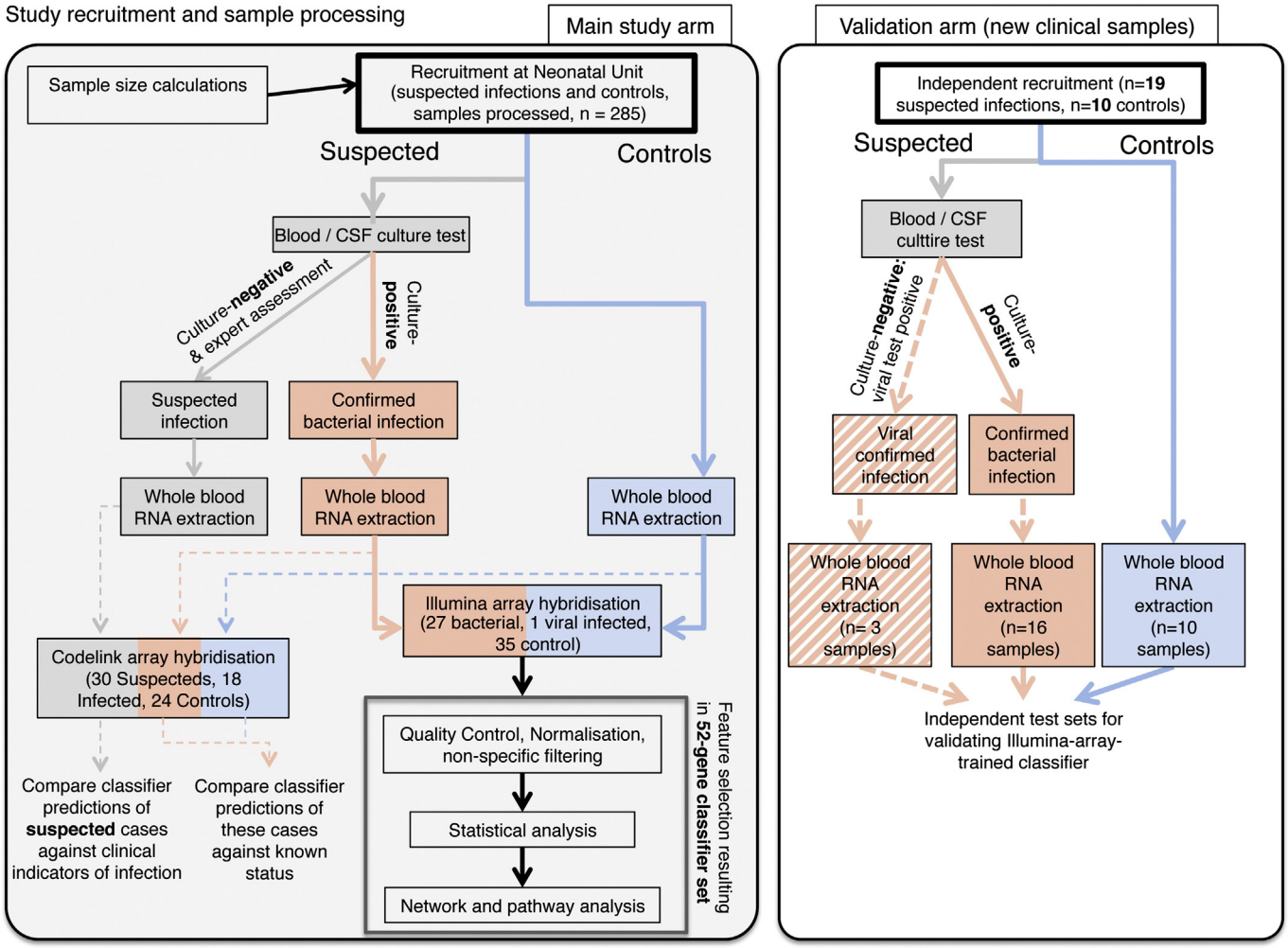

Fig. 1. Study recruitment and sample processing. This flow diagram depicts process of neonatal subject recruitment over sample processing and microarray hybridization. Boxes and arrows are color-coded as follows. Healthy (presenting for clinical reasons other than suspected infection) control neonate samples = blue; neonate samples of suspected but unconfirmed infections = gray; neonate samples with blood-culture test confirmed infection = pink; neonate samples with blood-culture negative test but confirmed viral infection = striped pink. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article. Figure 1 was adapted from Supplementary Figure 9 of Smith et al. 2014 [2] by permission from Macmillan Publishers Ltd: Nature Communications [2], copyright (2014).

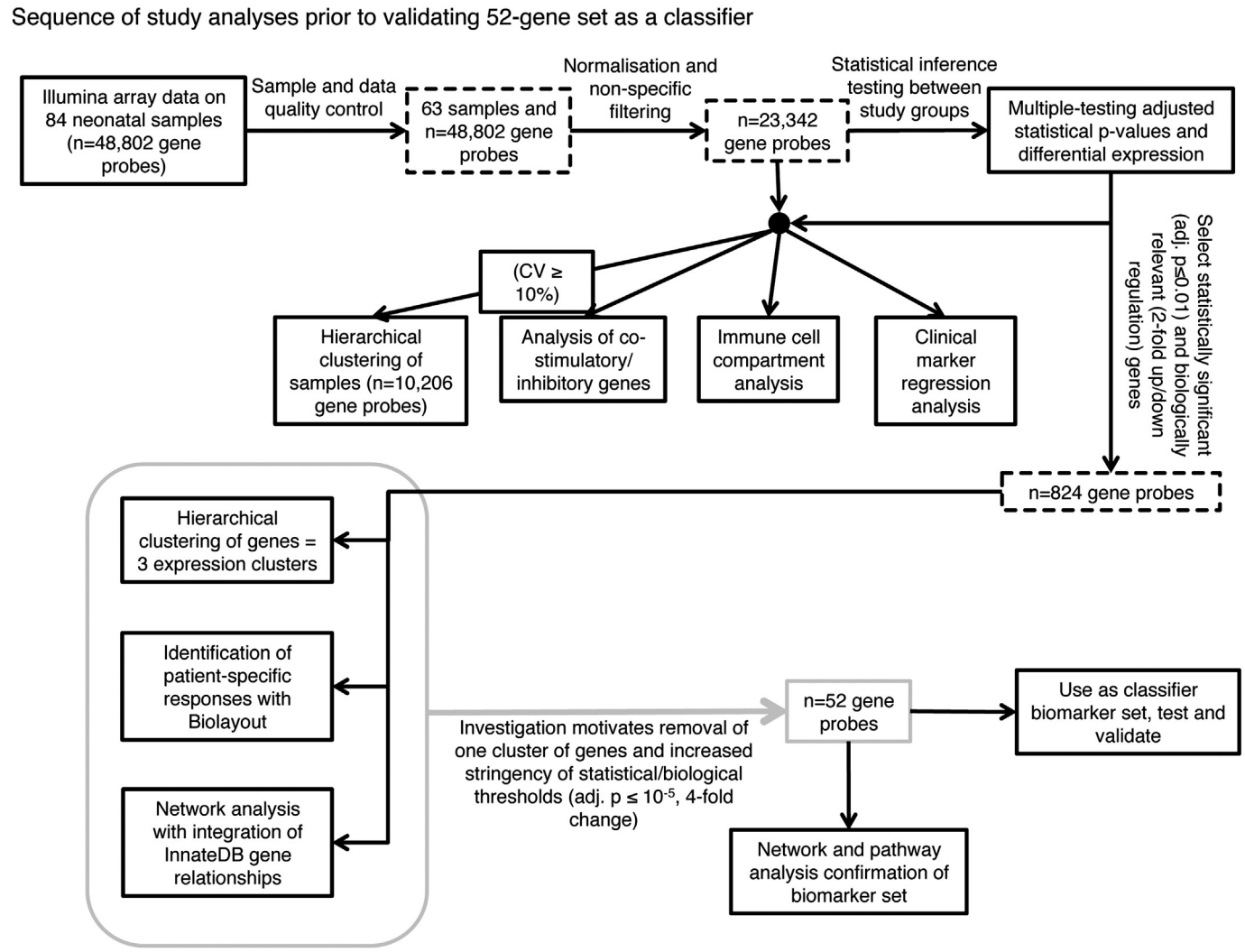

Fig. 2. Sequence of study analyses prior to validating 52-gene set as a classifier. This flow diagram identifies the sequence of analyses carried out on Illumina microarray data. The gray box indicates that the analyses within are used in combination to inform a subsequent result. Figure 2 was adapted from Supplementary Figure 10 of Smith et al. 2014 [2] by permission from Macmillan Publishers Ltd: Nature Communications [2], copyright (2014).

images were extracted using GenomeStudio (2010.3) Gene Expression Module(1.8.0) software.

For CodeLink arrays the biotin-labeled cRNA target is prepared by a linear amplification method using tailed oligo dT priming of total RNA. After second-strand cDNA synthesis, the cDNA undergoes an in vitro transcription (IVT) reaction to produce the target cRNA. Various quality control procedures are incorporated. Hybridization is performed overnight and post-hybridization processing includes a stringent wash to remove unbound and non-specifically hybridized target molecules and staining with Cy™5-streptavidin conjugate. Several non-stringent washes remove unbound conjugate. The bioarrays are then dried and scanned on the Agilent G2567A scanner at 5 nm resolution. Raw data were obtained from the scanned images using CodeLink™ EXPv4.1 (GE Healthcare) feature extraction software.

For Affymetrix HG-U133 plus 2.0 arrays biotin labeled cRNA target is prepared by a linear amplification method following reverse transcription of total RNA into T7 tailed double stranded cDNA. Biotinylated target cRNA was purified using RNeasy columns according to the manufacturer's instructions (QIAgen Ltd., Crawley, UK) and quantified by spectrophotometry. Fifteen micrograms of purified biotinylated cRNA was fragmented by heating for 35 min at 94 °C in the presence of magnesium ions, spiked with eukaryotic hybridization control and hybridized to HG-U133 plus 2.0 microarrays overnight at 45 °C. After hybridization the arrays were washed, stained with phycoerythrin coupled streptavidin and processed on the Affymetrix GeneChip Fluidics Workstation 400 using the EukGE-Ws2v4 protocol. Microarrays were then scanned using the Affymetrix GeneChip Scanner 3000 using GeneChip Operating Software instrument control and data acquisition system.

For Affymetrix U219 arrays total RNA was reverse transcribed to synthesize first-strand cDNA. This cDNA was then converted into a double-stranded DNA template for *in vitro* transcription to synthesize cRNA incorporating a biotin-conjugated nucleotide. This cRNA was then purified to remove unincorporated NTPs, salts, enzymes, and inorganic phosphate. The biotin-labeled cRNA was then fragmented and prepared for hybridization using the GeneChip HT Hybridization, Wash and Stain Kit for GeneTian (Affymetrix). Arrays were then processed and scanned on the Affymetrix GeneTitan Instrument as detailed in the Affymetrix GeneChip Command Console 2.0 User Guide.

Data normalization and analysis

For the computational and statistical pathway biology aspects of this study, a summary of the data analysis workflow is shown in Fig. 2. The

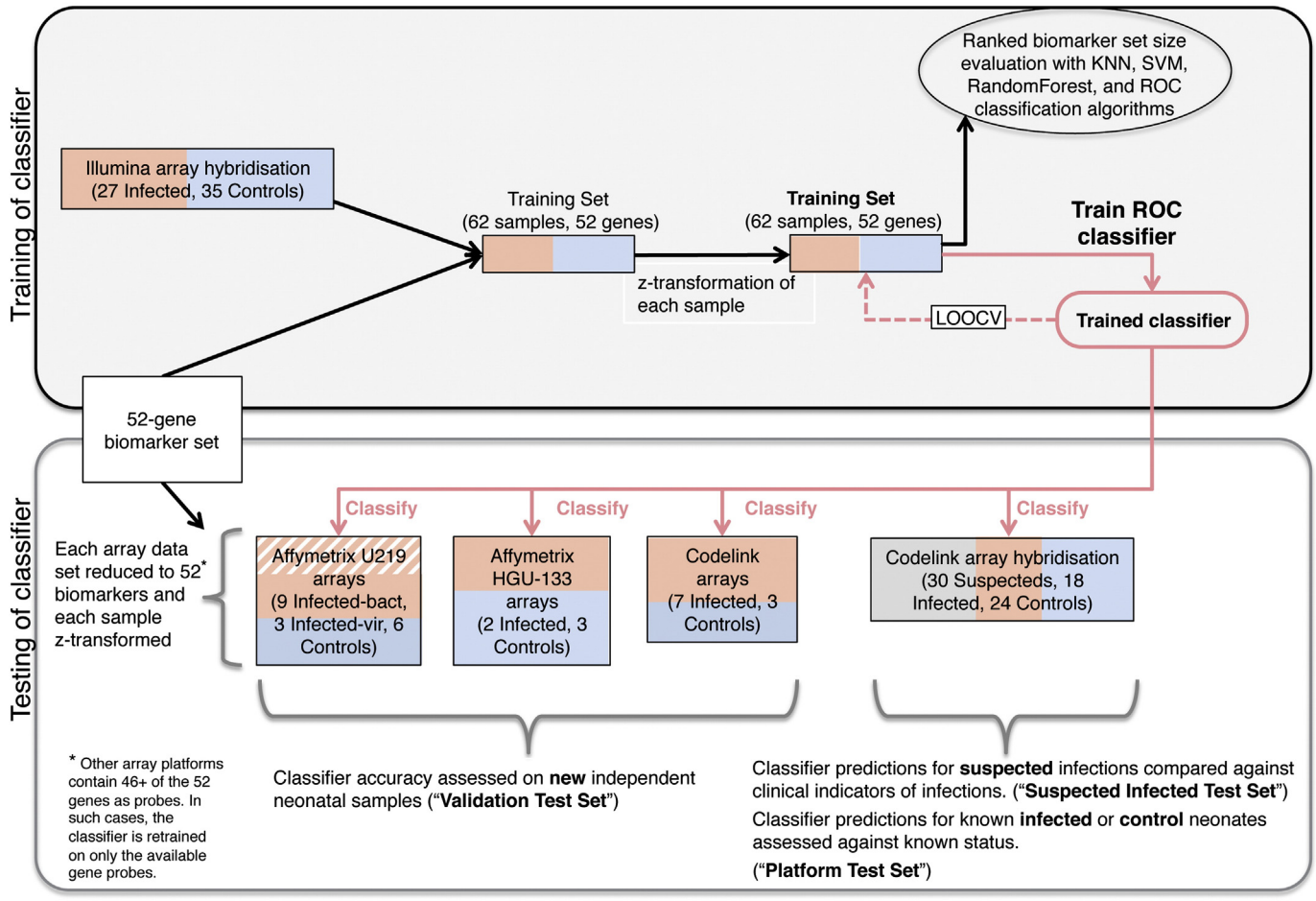

Fig. 3. Training and testing of 52-gene classifier of sepsis in neonates. This diagram details the stages comprising training and testing of the ROC-based classifier. Top box represents processes in the training of the classifier; bottom box represents processes in the testing of the classifier on various types of test sets. LOOCV stands for leave-one-out-cross-validation, which is the iterative process in which a single sample of the training set is predicted based on the classifier trained on all remaining samples. Black arrows are data processing steps; red arrows indicate classifier training and prediction steps. Sample color coding: healthy (presenting for other clinical reasons than suspected infection) control neonate samples = blue; neonate samples of suspected but unconfirmed infections = gray; neonate samples with blood-culture test confirmed infection = pink; neonate samples with blood-culture negative test but confirmed viral infection = striped pink. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article. Figure 3 was adapted from Supplementary Figure 11 of Smith et al. 2014 [2] by permission from Macmillan Publishers Ltd: Nature Communications [2], copyright (2014).

chronological processing stages cover: data quality control, processing, statistical analysis, gene feature selection and classifier testing and validation.

Data quality control: High-quality RNA (RNA integrity number (RIN) greater than 7) from infected and control infants were hybridized onto Illumina Human Whole-Genome Expression BeadChip HT-12 v3 microarrays comprising 48,802 features (human gene probes). Gene expression levels, distributions and controls were assessed using the arrayQualityMetrics package in Bioconductor [3]. A gender check was performed using Y-chromosome-specific loci.

Processing: Using the 'lumi' Bioconductor package, raw data from 63 samples were transformed using a variance stabilizing transformation before robust spline normalization to remove systematic between-sample variation. Microarray features that were not detected (using function 'detectionCall') on any of the arrays were removed from analysis and the remaining 23,342 features were used for subsequent statistical analysis.

Statistical analysis: Data were statistically examined to assess gestational age as a confounding factor. Within each sample group (control, infected), samples were age classified into bins based on the 33% and 66% corrected gestational age quantile values, yielding three age groupings. Per-gene hypotheses of differential expression between bacterial infection cases and control neonates were tested through linear modeling of the $\log_2$ scale expression values between groups and subsequent empirical Bayesian approaches to moderate the test statistic by pooling variance information from multiple genes (Bioconductor package 'limma' [4]). This included vertical *p*-value adjustment for multiple testing (Benjamini–Hochberg) to control for false discovery rate at a 1% level.

Gene feature selection for classifier: Computational network-based approaches were used to examine relationships in the data using correlation of gene expression and biological relationships. Statistically significant differentially expressed genes were examined further: heat maps and line graphs with hierarchical clustering by Euclidean distance were examined using Partek Genomics Suite v6.5, and visualization of networks of genes looking for patient-specific responses using BioLayout Express 3D [5]. These analyses were carried out step-wise using a pathway-biology approach, becoming more focused until a defined sub-network of 52 differentially expressed genes was identified [2]. The selected genes had adjusted *p* values of $\leq 10^{-5}$, fold changes of ≥ 4 and were highly connected in terms of biological pathways and networks.

Classifier training and testing: First, a simulation model based on these 52 genes was established to assess the relationship between the number of gene predictors and classification error and establish suitability of this gene set for use with a panel of classifier algorithms. This approach used leave-one-out cross-validation with four different machine learning methods: Random Forests, Support Vector Machines, K Nearest Neighbour, and ROC-based [6-9] (Fig. 3). Leave-one-out cross validation was repeated 100 times for each set of selected genes

following a random ordering of the data at each replication to minimize variability of the error estimates [10].

Next, the ROC-based classification method [9], (that does not require tuning of parameters and simplifies classification to a univariate decision that can easily be applied to independent data) was trained on the training set to learn the gene expression level differences that distinguish between controls and cases of infection. Internal accuracy of this classifier was tested through leave-one-out cross-validation on the training set, prior to its testing on independent data. Using the platform test set, a subset of 42 of the training set samples (18 infected, 24 controls) hybridized to CodeLink™ Whole Human Genome arrays, the trained ROC classifier was tested for platform-dependent performance. Subsequently, the classifier was tested for performance on completely new and independent neonatal samples, consisting of a further 26 samples (16 bacterially infected samples from 15 infants and ten control samples) which were run on CodeLink™ (seven infected, three control), Affymetrix® HG-U133 Plus 2.0 (two infected, three control) or Affymetrix® Human Genome U219 (nine infected, six control) arrays. Finally, the classifier was tested on $n = 30$ (hybridized to CodeLink arrays) new and independent cases where infection was suspected but not confirmed through blood culture and performance was compared against 'expert' clinical assessment (Table 2).

Discussion

We describe in this paper our detailed technical and analysis methodology for our data set describing the host response to neonatal infection. This data set is a unique repository of data describing the host response at the first sign of neonatal infection and has allowed identification of a 52-gene classifier that predicts bacterial infection with high accuracy. This data set lays the foundation for advancing diagnostic, prognostic and therapeutic strategies for neonatal sepsis and we hope will be of great value for future further investigations by the wider research community.

Conflict of interest

The authors declare there are no conflicting interests.

Acknowledgments

The authors would like to thank the infants and their parents for their participation in the study. This work was supported by the Wellcome Trust (WT066784) program grant, EU FP7 IAPP project ClouDx-i, Chief Scientists Office (ETM202) and BBSRC (BB/D019621/1) the Centre for Synthetic and Systems Biology at Edinburgh (SynthSys) supported by the BBSRC and EPSRC (BB/D019621/1) to P.G. and P.D.; MRC (G0701291) to K.L.F., P.D. and P.G. Teagasc (RMIS6018) funded D.J.L.'s participation in this study.

References

[1] C.L. Smith, P. Dickinson, T. Forster, M. Khondoker, M. Craigon, A. Ross, P. Storm, S. Burgess, P. Lacaze, B.J. Stenson, et al., Quantitative assessment of human whole blood RNA as a potential biomarker for infectious disease. Analyst 132 (2007) 1200–1209.
[2] C.L. Smith, P. Dickinson, T. Forster, M. Craigon, A. Ross, M.R. Khondoker, R. France, A. Ivens, D.J. Lynn, J. Orme, et al., Identification of a human neonatal immune-metabolic network associated with bacterial infection. Nat. Commun. 5 (2014) 4649.
[3] A. Kauffmann, R. Gentleman, W. Huber, arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics 25 (2009) 415–416.
[4] G.K. Smyth, Statistical Applications in Genetics and Molecular Biology Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microar. Stat. Appl. Genet. Mol. Biol. 3 (1) (2004) 1–25 Article 3.
[5] A. Theocharidis, S. van Dongen, A.J. Enright, T.C. Freeman, Network visualization and analysis of gene expression data using BioLayout Express(3D). Nat. Protoc. 4 (2009) 1535–1550.
[6] L. Breiman, Random Forests. Mach. Learn. 45 (1) (2001) 5–32.
[7] C. Cortes, V. Vapnik, Support-Vector Networks. Mach. Learn. 20 (3) (1995) 273–297.
[8] N.S. Altman, An Introduction to Kernel and Nearest-Neighbour Nonparametric Regression. Am. Stat. 46 (3) (1992) 175–185.
[9] M. Lauss, A. Frigyesi, T. Ryden, M. Hoglund, Robust assignment of cancer subtypes from expression data using a uni-variate gene expression average as classifier. BMC Cancer 10 (2010) 532.
[10] M.R. Khondoker, T.T. Bachmann, M. Mewissen, P. Dickinson, B. Dobrzelecki, C.J. Campbell, A.R. Mount, A.J. Walton, J. Crain, H. Schulze, et al., Multi-Factorial Analysis of Class Prediction Error: Estimating Optimal Number of Biomarkers for Various Classification Rules. J. Bioinf. Comp. Biol. 8 (6) (2010) 945–965.